

IDENTIFY POINTING AND WAVING IN GESTURE BASED HUMAN-MACHINE-INTERACTION

Dipl.-Ing. Tobias Nowack, M.Sc. Christoph Dutschmann

Technische Universität Ilmenau
Department of Mechanical Engineering
Biomechatronics Group

ABSTRACT

The gestures "pointing" and "waving" are investigated with regard to their characteristic properties and their suitability for using in gesture based Human-Machine-Interaction (HMI). The MS Kinect 2[®] as usable motion capturing device is proofed. A description of different basic types of gestures in human-to-human interaction is given and the requirements for the HMI are discussed. A general phase model in combination with the movements of a gesture execution is explained. For the technical recognition of these movement sequences, parameters are defined which are based on the joint data of the Kinect. Detecting the pointing gesture an angle, its angular velocity and optional a holding time are used. The waving gesture is also detected with the help of an angle and its periodicity. To evaluate these angles and the necessary threshold values, several experiments had been done.

Index Terms – human-machine-interface, gesture recognition, Microsoft Kinect 2 TM, pointing / waving

1. INTRODUCTION

Since 2010 affordable motion capturing devices (e.g. MS Kinect) are available. At the beginning, they have mainly been developed for gaming purposes. But short after introduction to the market, the systems also have been established for Human-Machine-Interaction (HMI). One of the first commercial deployments has been a system from Fraunhofer IOSB and BMW [1]. They have used the Kinect sensor to detect, when a worker points to a failure on a painted surface.

In case of HMI use of gestures, it is necessary to separate the different human gestures in purpose of its behavior. As "pointing" is used to localize any object during human to human communication other gestures are used to command other people in the context of their actual task. "Waving" is decoded as "hello" or "good bye" in most human to human interaction cases, but it can also be decoded to give the command "come over". For example the purpose of a driving task is to enable the driver continue moving the car until waving stops or change to a stop gesture.

2. GESTURES USES IN HUMAN-TO-(HUMAN / MACHINE) COMMUNICATION

If you have a closer look to human gestures, in general you will recognize that the human has different opportunities to support the gesture with additional information and to recognize its behavior. In human-to-human interaction, gestures will be mostly used additional, speech attended. Therefore, the main information will be communicated by the speech and the gesture only support additional information, which could be transported on these information channel more easily. For example if one communication partner will locate a special object, both could

already see, he will use the pointing gesture for the first approach. The attended speech will specify the object by additional parameters like color or size. “Please could you give me the blue box, right over there.”

Another reason to use gestures is depending on the situation. When both communication partners are limited in their possibility to use speech for communication, for example because there is a window in between them or there is too much noise in the surrounding, then humans also use gestures in communication. The aim in this case is primary not to support speech, the goal is to enable a basic communication mostly with a little set of commands. Sometimes there will a kind of predefined sign language be used e.g. the sign language of divers (see Figure 1).

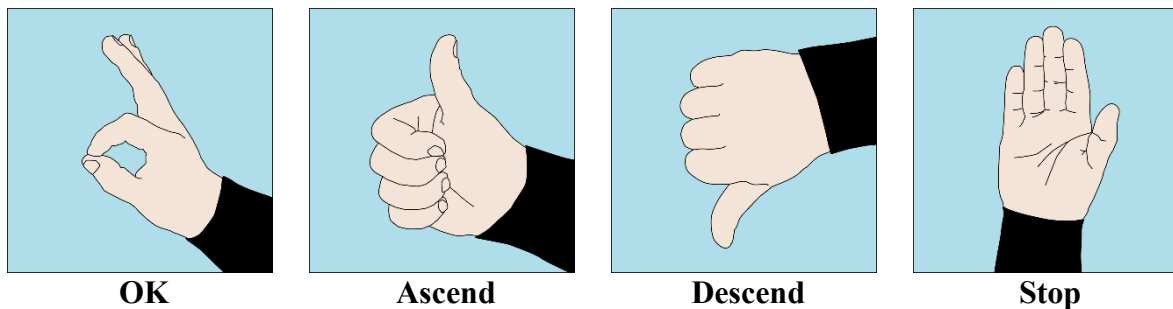


Figure 1: Diving Signs [2]

In human communications the knowledge of both communication partners are nearly the same. Wundt [3] studied the behavior of babies. He detected that at the age of approximately twelve months the baby points with the plan to communicate. The baby points towards a selected object and in a second step to an adult. Therefore, the baby will motivate the adult to have a closer look towards the selected object.

In case of HMI these early trained skills are already missing. Also the additional information like speech will not be necessarily included in the portfolio of sensors and software installed in an appropriate technical system. The knowledge of normal human behavior will not be implemented as well.

Humans will use the same additional knowledge if they see anybody waving. They will ask themselves which kind of information should be transported by this gesture. Depending on the actual situation and depending on the way the communication partner oriented his hand, there will be a different behavior of reaction. As described during the introduction waving can be decoded to say “Hello, here I am, have a look to me” or as “bye bye” over long distance as well as a moving command like “come over” or “continue moving”. It is same like pointing. By using the additional knowledge humans are able to interpret the situation quite well, whereas machines need additional rules or sensors.

3. TECHNICAL SENSORS

To use gesture recognition for controlling technical systems, it will be necessary to have systems for motion analysis. If you use a gesture recognition system to control technical systems in the field, it is additional necessary that these systems will be cheap, easy to use and without a big static setup. The Kinect System by Microsoft will fulfil the described requirements. With about 200 € it is cheap enough to mount it on a field system like a mobile robot. Designed to use for gaming it must be easy to use in any environments and without wearing additional clothes, markers or sensors.

The Kinect 2 sensor working at the USB 3 port provides a framerate of approximately 30 images per second and the Microsoft Library can detect up to 25 skeleton joints (see figure 2).

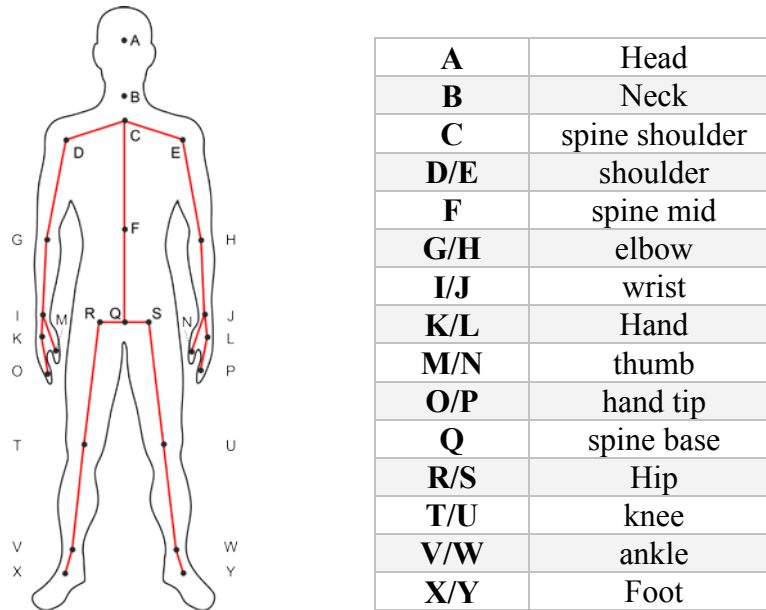


Figure 2: Stick figure with skeleton joints from Kinect 2®

4. PHASES OF GESTURE RECOGNITION

Kendon [4] and McNeill [5] had a closer look to the human communication and the use of gestures to support their speech. McNeill [5] identifies four (five) types of gestures which will be used for communication. There are iconics, metaphors, beat gestures, deictics (and cohesives). For the daily use in speech attendance only the first four types might be interesting. To fulfill the needs in HMI only the deictic and some iconic might be really useful. The beat gestures are important if you prefer to keep the speed of your speech, special if you use the speed as characteristic feature of the speech.

Metaphoric gestures need the basic knowledge of the communication partners. Using the communication model of Shannon and Weaver [6] (see Figure 3) both communication partners can decode the message (command) if the code is the same for coding and decoding.

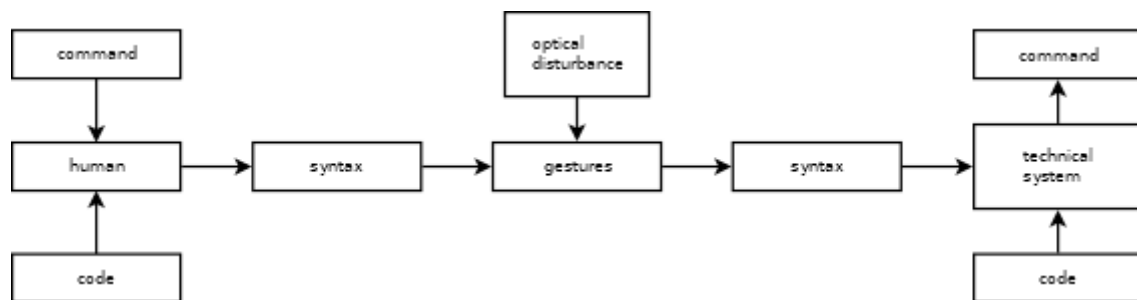


Figure 3: Communication with gestures, enhanced model based on Shannon and Weaver [6] and Herczeg [7]

The problem of human coding a command into a gesture and to decode the gesture back into the right command on the side of the technical system will be similar for the other two classes. For iconic gestures instead of metaphoric gestures, the way of performing these gestures is more formalized. So it is much easier to decode this version because there are defined rules which hopefully could be implemented into a technical system. There will be a closer look to the waving gesture in chapter 6.2.

The class of deictic gestures is the smallest one, because there is only a summarization of the different forms of pointing. Having a closer look to that class, the different forms are as similar as they could be described with an easy algorithm see chapter 6.1.

Within his studies McNeill identified that there is a standard sequence which describes gestures in general. His model contains that three phases of movements are nearly compulsory and considered necessary (preparation, stroke and retraction) whereas the two optional holds only “be held more or less briefly [...] when, for some reason, the stroke onset is delayed” [5] or the meaning of the stroke will be reinforced. Preparation and retraction in the speech attended use of gestures are optional, if the orator will perform one gesture directly after the one before.

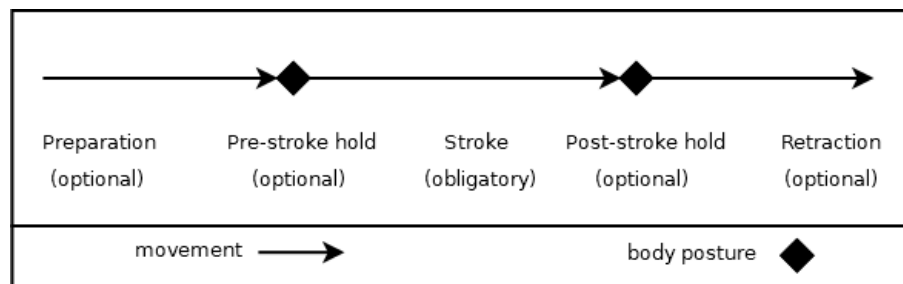


Figure 4: Phase Model from McNeill, [5]
picture [8]

5. TECHNICAL PHASE DESCRIPTION

For a gesture description which could be decoded with the Microsoft Kinect 2 sensor it is necessary to identify several parameters which describes the individual movement from humans performing gestures. The humans have different cultural imprinting, different educational background and perform the gesture like they think, this gesture would present the command in the best and precise way.

Specific environmental conditions could not get described in advance, because of the scenario “mobile HMI”. In this case, the parameters for the gesture description should be extracted from the joint data. Even the absolute values of the joint data depends on the actual setup. The joint data are points within 3D. The x and y coordinates are presented in px, starting in the left upper corner of the picture, whereas the z component presents the distance between the Kinect Sensor and the visual plane in cm.

Use an angle between vectors defined by the joints seems to be the simplest way. The vectors do not have to correspond with natural human bones, they are only a mathematic trick to reduce the complexity. As you can see in figure 5 special threshold values of the angle Φ or the angular velocity $\dot{\Phi}$ in combination with a holding time t define the whole gesture cycle. The static body postures are primarily defined by Φ and t , the phase transition by changing $\dot{\Phi}$ and the dynamic phases by a combination of Φ and $\dot{\Phi}$.

With this general phase model natural pointing and waving has been analyzed.

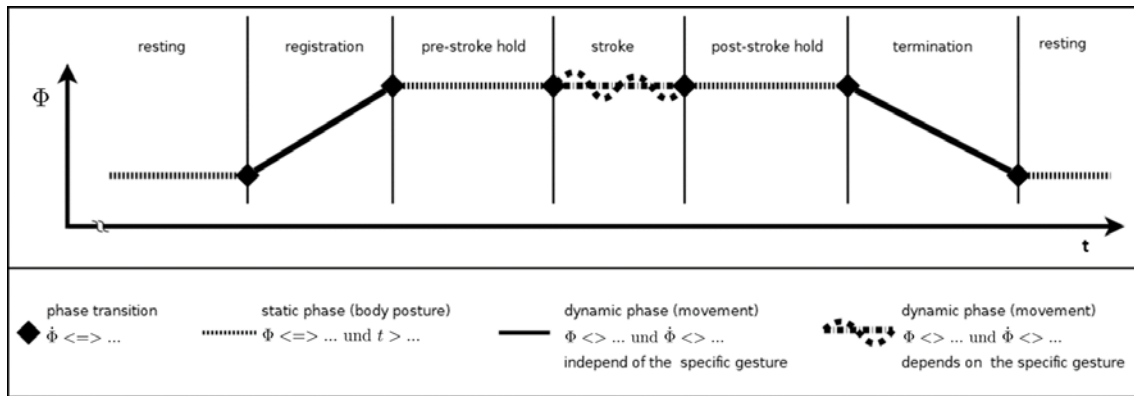


Figure 5: Full phase model of technical gesture description [8]

6. PARAMETER DEFINITION OF GESTURES

To evaluate the complete parameter set, consisting of an angle describing the gesture and the threshold values, several experiments had been done. All experiments had been done with a recording software shown in figure 6. This software give the investigator the opportunity to separate the whole recording in small sequences. Every sequence contains the individual performed gesture for a given command. The commands have been differ a bit during the groups even between the experiments. The aim of the experimental design was to collect as much different versions, how humans translate a command into a gesture.



Figure 6: Recording software for experiments with stick figure and real picture [9]

To optimize and evaluate the threshold values of the detection algorithms the detection rate have been used. Because there have been two items with two properties each, a 2 x 2 table has been used to calculate the detection rate over all. Depending on which gesture should be investigated there have been these two features:

- 1.) Experimentee performed investigated gesture, which must be detected by the algorithm
- 2.) Experimentee performed a control gesture, which must be declined by the algorithm

The result of the algorithm will be the second item with these two features:

- 1) Investigated gesture selected
- 2) Gesture declined as not performing the searched command.

6.1 Pointing gesture

Which angle will be the most relevant angle for pointing? During the development there have been identified two angles which are relevant to describe the body posture, most people would perform as pointing, if they know, that the system cannot identify finger and hand correctly. The angles are the elbow angle and the shoulder angle.

Pointing means the arm is stretched (elbow angle $> 160^\circ$) and the arm is lifted from the resting position (shoulder angle $> 20^\circ$). This body posture has to be hold for about 0.5 second [10]. This algorithm could not detect if somebody is pointing with a flexed arm and directed with the index finger.

In the theory it is explained that only one angle, its velocity and optional a holding time is necessary to describe the gesture. The angle must not necessarily exist as a real joint. So the result of the optimization have been the angle between a reference vector (Torso) along the spine (spine base – spine shoulder, \bar{T}) and a virtual arm (hand up to shoulder, \overline{HS}).

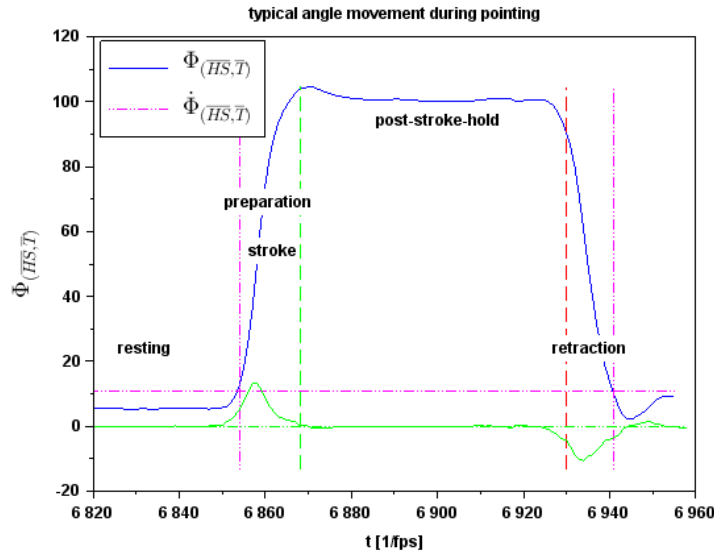


Figure 7: typical angle movement during pointing

Figure 7 shows the typical angle movement over the time for a pointing gesture. The experimentees always starts in a resting position, hands by side of the body. Because the pointing information is transported with the body posture during the post-stroke-hold, for pointing the preparation and the stroke phase will be combined in a short movement direct starting from the resting position. The post-stroke-hold is characterized by an immobility during the holding time. Immobility in case of a pointing arm means that the angle velocity is below $1^\circ/\text{frame}$.

$$|\dot{\Phi}_{(\overline{HS}, \bar{T})}| \leq \frac{1^\circ}{\text{frame}}$$

requiring $t \geq 6 \text{ frames}$

Equation 1: immobility during post-stroke-hold

This immobility must hold over at least 6 frames. At the end of the post-stroke-hold the retraction to the resting position starts. If another movement starts after the algorithm detects a post-stroke-hold it would not be a post-stroke-hold. In some cases it might be a pre-stroke-hold before the stroke phase of another gesture will be performed. The retraction phase for the pointing gesture is described as a continuously movement back to the resting position, so the angle velocity $\dot{\Phi}$ will be less than 0 °/frame.

$$\dot{\Phi}_{(\overline{HS}, \overline{T})} \leq \frac{0^\circ}{frame}$$

until $\Phi_{(\overline{HS}, \overline{T})} < 21^\circ$

Equation 2: movement during retraction phase (pointing gesture)

In the experiments with 15 participants (8 female, 7 male, mainly students) this algorithm has an accuracy of 85.7 % when the experimentees will be positioned in front of the sensor with a maximum disorientation of 30° (angle between shoulder line and the line of sight). The accuracy to detect pointing is 81.5% but the accuracy to decline pointing is up to 93.5%.

6.2 Waving gesture

On the other hand waving would have been a good partner to proof the basic algorithm. Instead of pointing, for waving the main information is coded by the movement of the arm not by the final static position. We can call the waving gesture a dynamic gesture whereas the pointing gesture will be called a static gesture. The question was whether the basic model of technical gesture recognition, using one calculated angle, based on the detected joints, and several threshold values of this angle would be also enough for dynamic gestures.

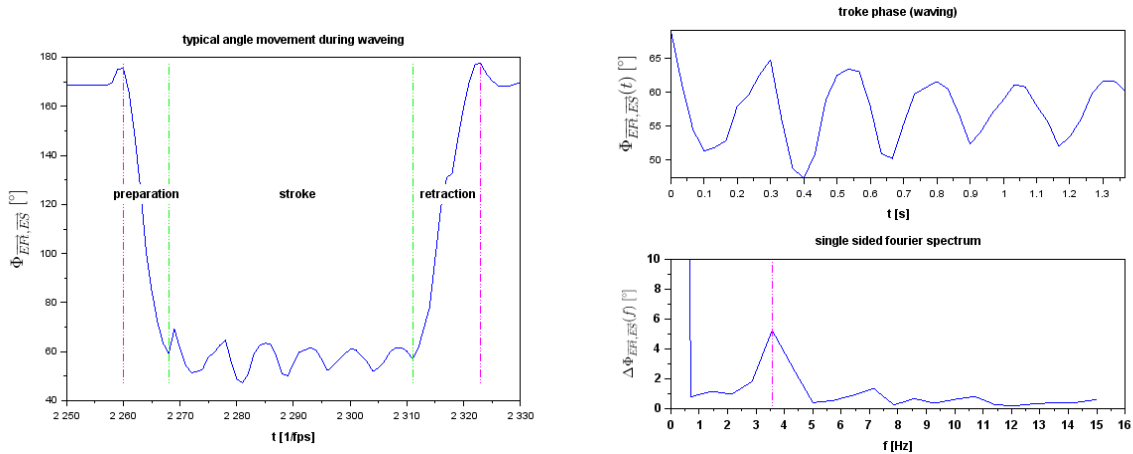


Figure 8: a) typical angle movement during waving;
b) cutted stroke sequence and single sided fourier spectrum

Figure 8a shows, that the typical angle movement during waving has a nearly similar shape than the version for pointing. The best angle to describe waving is an angle between the vector fingertip – elbow (\overline{EFt}) and elbow – shoulder (\overline{ES}). In the resting position, the arms are beside the body, that means the elbow is stretched (more than 160°). During the preparation phase the experimentees flex the elbow up to less than 138° and start with repetitive flexing the elbow.

This repetitive flexing characterizes the waving and could be described by the frequency of the alternating signal. As shown in figure 8 b) this alternating signal will have a frequency between 0.95 and 6.6 Hz with a minimum alternating angle $\Delta\Phi_{(\overline{Eft}, \overline{ES})}$ more than 3.3° .

$$0.95 < f < 6.6 \text{ Hz}$$

$$\text{with } \Delta\Phi_{(\overline{Eft}, \overline{ES})} > 3.3^\circ$$

Equation 3: threshold values to characterize stroke phase during waving

An experiment with 13 experimentees (9 female, 4 male, mainly students) proved this algorithm with a detection rate of 89.0 % for waving in term to say “good bye”.

In case that waving could have two different meanings, the experimentees used two extremely different gestures. Waving to say good bye had been performed as a movement from the left to the right whereas waving to command come over had been performed as a movement forwards and backwards. When the command was only “waving” without any further information every experimentee performed the waving to say “good bye” gesture.

The detection rate will decrease to 68.3 % when especially the case of waving to command “come over” will be detected by this algorithm.

7. Discussion

As described the experimental design should guarantee that every experimentee could perform the gesture for a given command like he would do it in human-human-communication. With this idea experimentees did not get any information about the parameter for calculating the gesture. So there are various interpretations of some commands.

The pointing algorithm separates whether the experimentee points with the right or the left arm. But for example sometimes an experimentee points also with both hands (see figure 9 a). If a technical system will use the pointing information to perform an action it must be clear which information is correct. So it might be necessary to identify the target. On the other hand it could be necessary to have a closer look to the two forms of double pointing. Both pointing vectors are oriented in the same direction (see Figure 9 a), or both pointing vectors are oriented in separate directions but performed within a sequence. The first version is used when the human will enhance his command. The version with pointing in two directions within a sequence could be interpreted like a command sequence: the object and the target.

For the waving gesture the problem with the two different meanings has been discussed already.

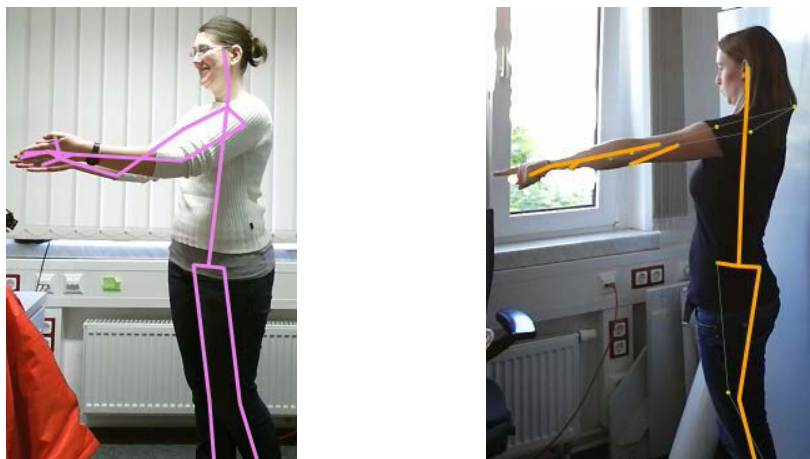


Figure 9: a) pointing with both hands in the same direction b) problems with detecting all 25 joints

For the waving gesture as well as for the pointing gesture the problem of detecting the joints by the Kinect has to be focused. The Microsoft library can detect joints as long as the body part is in the field of view. When the interesting body part will be covered, for example by other body parts (see figure 9 b), the joints might be interpolated. For the pointing gesture interpolation is only a problem if the user will be turned more than 60° from the line of sight. For the dynamic movement of waving the covering of joints might be a systematic problem. During the experiments only about 1.2 % of the relevant joints have been covered. The longest continuous sequence of covered data has been 1.5 s (47 frames). This period was not during a waving sequence.

Another big problem with these algorithms is, that they both need the whole sequence to detect whether it is pointing or waving. As Herczeg [7] explains, the acceptance of interactive computer systems is dominated by the response time. As long as the whole gesture sequence will be needed to determine the meaning the response time will be higher. The human will not have an opportunity to influence the detection when the system will identify that the detection will probably fail. To increase the speed, it must be checked if both algorithms can work together in a twostep system.

The last question which has to be discussed is the question of the responsible user. In the experimental design there have been in the laboratory: an experimentee, the investigator and an experiment leader. Only the experimentee has been located in the field of view by the Kinect. In that case it is clear that only this human can perform a gesture. But if the algorithm will be used outside the laboratory, there will be more than one person in front of the Kinect. As the Kinect 2 could decode the movement of up to six persons the question is really difficult to answer. As long as only one user will perform a detectable gesture the question could clearly be answered. But if the algorithms will detect two gestures at the same time, how will these be interpreted by the technical system? The question of logical interpretation of gesture detection in case of interactive use of gestures has to be discussed in the near future.

8. Summary and outlook

Within this article it could be demonstrated, that with a single angle, the angle velocity and an optional holding time it is possible to describe gestures for technical detection by a motion analysis system. The whole gesture sequence consists of three necessary movement phases (preparation, stroke and retraction) and optional two additional holds (pre- and post-stroke hold). For the pointing gesture as static gesture and for the waving gesture as dynamic gesture this model could be proved with defined threshold values and Microsoft Kinect as motion analysis sensor.

In the next steps it would be necessary to identify more gestures by using this model. Then it would be possible to investigate how these different angles have to be combined to speed up the response time.

A system with two interaction steps for pointing and defining objects has already been realized. This system uses the pointing data for pre detecting objects which should be handled. These pre detected objects have to be confirmed by the user on a touch screen [8 and 10].

REFERENCES

- [1] A. Schick, and O. Sauer, “Gestenbasierte Qualitätskontrolle: Intuitive Mensch-Maschine-Interaktion in der Industrie,” <http://www.iosb.fraunhofer.de/servlet/is/44590/Artikel%20Gesteninteraktion.pdf>, 2013.
- [2] P. Southwood, “Dive hand signal,” de.wikipedia.org, 2011.
- [3] W. Wundt, Völkerpsychologie: eine Untersuchung der Entwicklungsgesetze von Sprache, Mythos, und Sitte, W. Engelmann, Leipzig 1900.
- [4] A. Kendon, Gesture: Visible action as utterance, Cambridge University Press, Cambridge, New York, 2004.
- [5] D. McNeill, Hand and mind: What gestures reveal about thought, University of Chicago Press, Chicago, 1995, ©1992.
- [6] C. E. Shannon, and W. Weaver, The mathematical theory of communication, University of Illinois Press, Urbana, 1964, ©1949.
- [7] M. Herczeg, Software-Ergonomie: Theorien, Modelle und Kriterien für gebrauchstaugliche interaktive Computersysteme. Lehrbuchreihe interaktive Medien, Oldenbourg, München, 3., vollständig überarbeitete und erweiterte Auflage, 2009.
- [8] T.F. Nowack, Mensch-Technik-Interaktion mittels Freiraumgesten, Ilmenau, 2017 (im print).
- [9] S. Wenzel, T. Nowack, and P. Kurtz, “Unterstützung der Körperhaltungsbewertung laut Leitmerkmalmethode ”Ziehen und Schieben” mit Hilfe einer Tiefenkamera,” In Arbeit in komplexen Systemen - digital, vernetzt, human?!, S. B.4.9. GfA-Press, Dortmund, 2016.
- [10] T. Nowack, S. Lutherdt, S. Jehring, et. Al., “Detecting Deictic Gestures for Control of Mobile Robots,” Advances in Human Factors in Robots and Unmanned Systems, P. Savage-Knepshield and J. Chen (eds.), DOI 10.1007/978-3-319-41959-6_8, Springer International Publishing Switzerland. 2017.

CONTACTS

Dipl.-Ing. Tobias Nowack
Ma.Sc. Christoph Dutschmann

tobias.nowack@tu-ilmenau.de
christoph.dutschmann@gmail.com